Multi-Way Tensor Decomposition of Joint Regulatory Architectures as a Competing Risk Predictor of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding the simultaneous interplay of genomic and epigenetic regulation is critical for predicting tumor lethality, yet traditional models often isolate molecular layers or confound disease-specific mortality with general host fitness. To address this challenge, we constructed a three-dimensional multi-omic tensor comprising gene expression, copy number alterations, and promoter methylation across 1,417 primary breast cancer samples. We applied multi-way Canonical Polyadic tensor decomposition to extract patient-level latent factors that holistically encode joint cross-layer regulatory architectures rather than isolated per-gene anomalies. The prognostic utility of these latent molecular states was then evaluated using a cause-specific competing risks framework to explicitly isolate disease-specific mortality from other-cause deaths, adjusting for standard clinicopathological variables. Our decomposition identified 11 structurally stable latent factors representing distinct regulatory modes, including coupled dosage-driven and epigenetic-driven axes, which demonstrated independence from established clinical taxonomies. Cause-specific Cox proportional hazards modeling revealed that specific joint regulatory states are powerful predictors of breast cancer lethality; for instance, a high-risk architectural state enriched for proliferation genes significantly increased disease-specific mortality risk, while other configurations conferred protective effects. Incorporating these tensor-derived factors improved the prognostic concordance index for disease-specific death from 0.6834 to 0.7072 without improving predictions for other-cause mortality, confirming their specificity to cancer lethality. Validated through a leave-one-cohort-out cross-validation strategy, this multi-omic tensor approach provides a reproducible and biologically grounded method for stratifying breast cancer lethality.

1 Introduction

Breast cancer is a highly heterogeneous disease driven by complex alterations across multiple molecular layers, including genomic copy number changes, epigenetic modifications, and transcriptomic dysregulation. Accurately predicting tumor lethality requires an understanding of the simultaneous interplay among these regulatory mechanisms. However, traditional prognostic models often isolate these molecular layers, analyzing gene expression, copy number alterations (CNA), and DNA methylation independently. When multi-omic integration is attempted, it frequently relies on sequentially degraded residual models or inappropriately strips away macroscopic aneuploidy signals. These approaches often fail to capture the holistic, joint regulatory architectures that drive cancer progression, instead focusing on noisy, isolated per-gene anomalies. Furthermore, many existing survival models confound disease-specific mortality with general host fitness or aging, limiting their ability to identify true biological drivers of cancer lethality.

To address these methodological and biological challenges, we propose a multi-way tensor decomposition approach to model the joint regulatory architecture of breast cancer. By constructing a three-dimensional multi-omic tensor encompassing samples, genes, and molecular layers, we can simultaneously evaluate the contributions of expression, CNA, and promoter methylation. We applied Canonical Polyadic (CP) tensor decomposition to extract patient-level latent factors that encode continuous, cross-layer molecular states. This unsupervised approach allows for the identification of coupled regulatory axes—such as joint dosage-driven or epigenetic-driven programs—without being constrained by established static taxonomies.

The prognostic utility of these latent molecular states must be evaluated in a manner that explicitly isolates tumor biology from non-disease mortality. To achieve this, we employed a cause-specific competing

risks framework, distinguishing disease-specific deaths from other-cause mortality. By systematically adjusting for standard clinicopathological variables, including age at diagnosis, tumor size, histological grade, and established molecular subtypes, we ensure that the tensor-derived factors capture novel biological mortality drivers rather than serving as redundant proxies for tumor burden or patient age.

In this study, we applied this framework to a cohort of 1,417 primary breast cancer samples. Our decomposition identified 11 structurally stable latent factors representing distinct regulatory modes. Cause-specific Cox proportional hazards modeling revealed that specific joint regulatory states are powerful, independent predictors of breast cancer lethality. For example, a high-risk architectural state enriched for proliferation genes significantly increased disease-specific mortality risk, while other configurations conferred protective effects. Incorporating these tensor-derived factors improved prognostic accuracy for disease-specific death without improving predictions for other-cause mortality, confirming their specificity to cancer lethality. Validated through a rigorous leave-one-cohort-out cross-validation strategy, this multi-omic tensor approach provides a reproducible and biologically grounded method for stratifying breast cancer lethality.

2 Methods

2.1 Cohort alignment and multi-omic tensor construction

To capture the simultaneous interplay of genomic and epigenetic regulation, we constructed a three-dimensional multi-omic tensor across 1,417 primary breast cancer samples with complete data for gene expression, copy number alterations (CNA), and promoter methylation. We first intersected the feature space across the three molecular layers. To prevent the introduction of artificial regulatory signals, we explicitly dropped any genes containing missing values—particularly from the methylation matrix, which exhibited approximately 3% missingness—rather than applying imputation techniques. From this complete-case intersection, we selected the top 3,000 most variable genes based on the expression matrix to maximize the signal-to-noise ratio and computational efficiency.

To prevent layers with larger numerical ranges from dominating the subsequent decomposition, each molecular matrix was independently normalized. For the continuous expression and methylation data, we applied a standard z-score normalization per gene across all samples to center the distributions and equalize variances. For the discrete CNA data (ranging from -2 to 2), we employed a robust scaling approach by dividing by the maximum absolute value per gene. This preserved the biological distinctions between copy number gain, loss, and neutral states without introducing continuous artifacts. The three aligned and normalized matrices were then stacked along a third axis to form a multi-omic tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, where $I = 1417$ (samples), $J = 3000$ (genes), and $K = 3$ (molecular layers).

2.2 Multi-way tensor decomposition and stability analysis

We applied Canonical Polyadic (CP) tensor decomposition, also known as PARAFAC, to the constructed 3D tensor to extract patient-level latent factors that holistically encode joint cross-layer regulatory architectures. The CP decomposition factorizes the tensor into a sum of rank-one tensors:

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \tag{1}$$

where R is the number of latent components (rank), $\circ$ denotes the vector outer product, and $\mathbf{a}_r \in \mathbb{R}^I$, $\mathbf{b}_r \in \mathbb{R}^J$, and $\mathbf{c}_r \in \mathbb{R}^K$ represent the sample-mode, gene-mode, and layer-mode factor vectors for the r-th component, respectively.

The decomposition was performed using the `tensorly` library. We evaluated a range of ranks ($R = 5$ to 30) by monitoring the reconstruction error and calculating the Core Consistency Diagnostic (CORCONDIA) to select the optimal number of latent components, which yielded an optimal rank of $R = 11$. To ensure the decomposition did not overfit to sample composition, we assessed the structural stability of the gene-mode weights across five distinct molecular accrual partitions (Cohorts 1 to 5). We performed CP decomposition on leave-one-cohort-out subsets and calculated the correlation of the gene-mode weight vectors between the

full-cohort model and the subset models, confirming that the extracted architectures represent biological constants rather than batch artifacts.

2.3 Biological grounding and clinical confounder analysis

To prove that the latent factors represent coherent biological states rather than redundant clinical proxies, we performed a feature importance analysis on the gene-mode weights. For the most prognostic factors, we extracted the top 50 genes with the highest absolute weights and performed Over-Representation Analysis (ORA) using standard hypergeometric tests. We evaluated the overlap of these top genes with a panel of 173 known cancer drivers and correlated their expression with established proliferation calls, explicitly excluding samples with missing proliferation data to avoid skewing the results.

Furthermore, we correlated the extracted sample-mode scores ($\mathbf{a}_r$) with key clinical variables, including age at diagnosis, tumor size, and histological grade. This redundancy analysis was critical to rule out the fallacy of misinterpreting indolent, early-stage disease or older age as a novel host fitness trait, ensuring the factors capture intrinsic molecular reorganization.

2.4 Competing risks framework and survival modeling

To explicitly isolate disease-specific mortality from other-cause deaths, we evaluated the prognostic utility of the latent molecular states using a cause-specific competing risks framework. We defined two separate binary events based on vital status and overall survival months: a disease-specific event (death from breast cancer) and an other-cause event.

Patients were stratified into high- and low-score groups using a median split on the sample-mode scores of the top factors. We estimated the cumulative incidence of both event types for these strata using the Aalen-Johansen estimator, visually assessing the separation of disease-specific mortality from non-disease mortality.

We then fit cause-specific Cox proportional hazards models for both event types. To rigorously adjust for established taxonomies, we defined three models:

1. **Baseline Model:** Adjusted for established integrative clusters and age at diagnosis.
2. **Expanded Baseline Model:** Added histological grade and tumor size.
3. **Full Model:** Included the Expanded Baseline covariates plus the continuous latent factor scores.

We reported the change in Hazard Ratios (HRs) for the latent factors before and after adding clinicopathological covariates. Model performance was quantitatively assessed by comparing the Concordance Index (C-index) between the Expanded Baseline and Full models, and by calculating Brier scores and generating calibration curves to evaluate predictive accuracy at 5- and 10-year time points.

2.5 Accrual cohort replication and cross-validation

Finally, we validated the predictive independence of our findings using a leave-one-cohort-out cross-validation strategy based on the five accrual cohorts. To prevent data leakage, layer-specific normalization was performed strictly within the training cohorts, and the resulting scaling parameters were applied to the held-out cohort. The held-out samples were then projected onto the training gene-mode and layer-mode factors to obtain independent sample scores:

$$\mathbf{A}_{\text{test}} = \mathcal{X}_{\text{test}} \times_2 (\mathbf{B}^{+}_{\text{train}}) \times_3 (\mathbf{C}^{+}_{\text{train}}) \tag{2}$$

where $\times_n$ denotes the n-mode tensor product and $+$ denotes the Moore-Penrose pseudoinverse. We re-ran the cause-specific Cox models on these independent sample scores to confirm the generalizability of the joint regulatory signals as predictors of breast cancer lethality.

3 Results

3.1 Multi-omic tensor decomposition and joint regulatory architecture discovery

To characterize the joint regulatory architecture of primary breast tumors, we applied multi-way Canonical Polyadic (CP) decomposition to a three-dimensional multi-omic tensor comprising 1,417 samples, 3,000 high-variance genes, and three molecular layers (gene expression, copy number alterations, and promoter methylation). We evaluated a range of tensor ranks to determine the optimal number of latent components. Based on the Core Consistency Diagnostic (CORCONDIA) score balanced against a steadily decreasing relative reconstruction error, an optimal rank of 11 was selected (). This rank allowed for the extraction of 11 latent factors (F1–F11) that encode holistic, cross-layer molecular states rather than isolated single-layer anomalies.

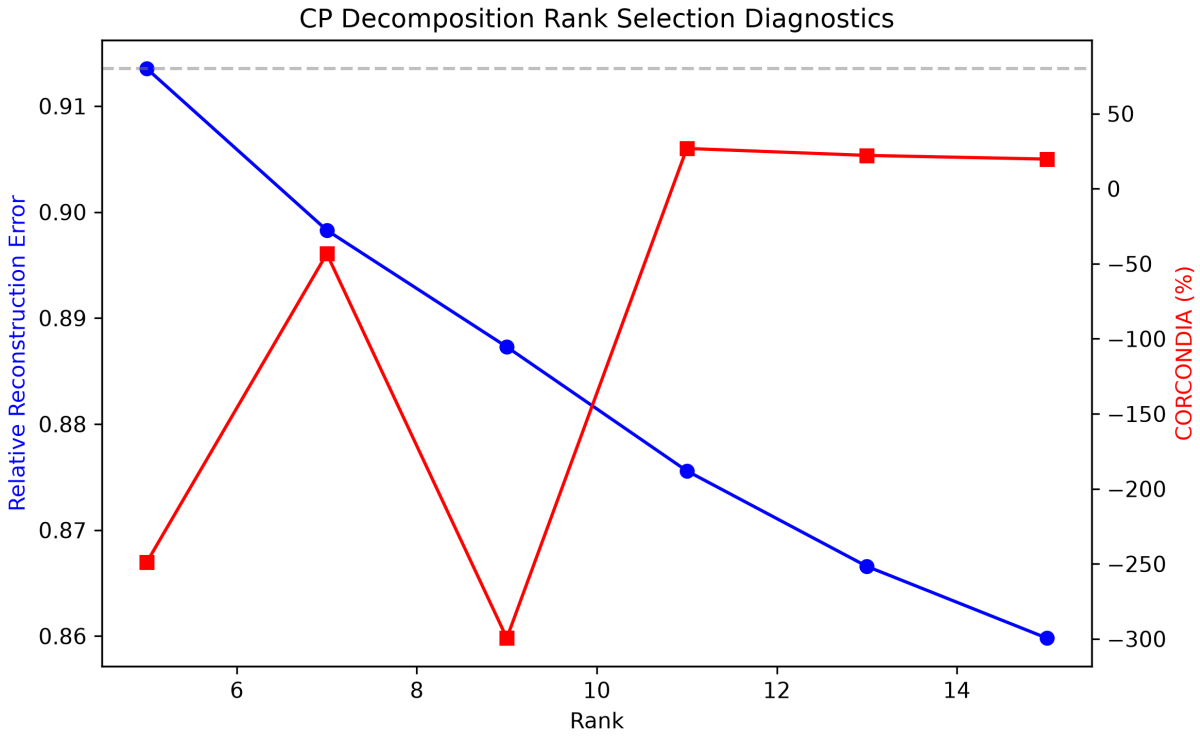


Figure 1: Canonical Polyadic (CP) decomposition rank selection diagnostics for the 3D multi-omics tensor. The plot displays the relative reconstruction error (blue, left axis) and CORCONDIA score in % (red, right axis) across varying tensor ranks. A rank of 11 was selected for the analysis, corresponding to a distinct peak in the CORCONDIA score balanced against a steadily decreasing reconstruction error. This optimal rank justifies the extraction of the 11 latent factors used to encode holistic cross-layer molecular states across gene expression, copy number alterations, and promoter methylation, which subsequently revealed highly prognostic joint regulatory architectures independent of standard clinicopathological variables.

To ensure that the extracted regulatory architectures represent biological invariants rather than batch-specific artifacts, we rigorously validated the structural stability of the factors using a leave-one-cohort-out (LOCO) approach across the five distinct accrual cohorts. The gene-mode factor weights maintained consistently high correlation across all five partitions (), confirming that the tensor decomposition successfully captured robust and reproducible molecular signals.

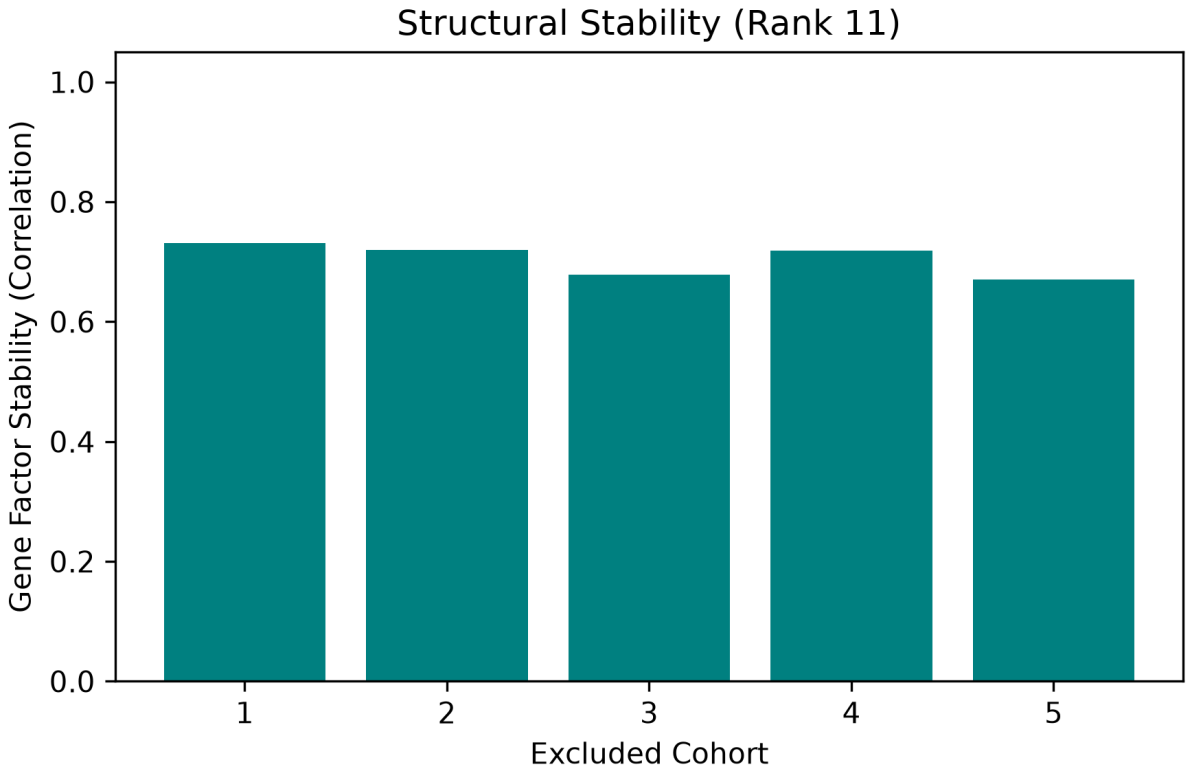


Figure 2: Structural stability of the 11 latent factors extracted via multi-way tensor decomposition, evaluated using a leave-one-cohort-out (LOCO) approach across five accrual cohorts. The bar chart displays the gene-mode factor stability, measured by correlation, when each specific cohort is excluded from the analysis. The consistently high correlation across all five partitions demonstrates that the gene-mode factor weights maintain high structural stability, confirming that the identified joint regulatory architectures represent robust biological invariants rather than batch-specific artifacts.

3.2 Independence from established clinical taxonomies

A critical requirement for novel prognostic biomarkers is their independence from existing clinical and molecular labels. While several of the 11 latent factors demonstrated significant associations with established molecular subtypes—for example, F3 and F4 showed strong associations with Integrative Clusters ($p < 10^{-20}$, Kruskal-Wallis test)—their sample-mode scores exhibited predominantly weak correlations with standard clinicopathological variables. As shown in , the latent factors are largely uncorrelated with age at diagnosis, tumor size, and histological grade. This independence suggests that the tensor-derived factors capture intrinsic molecular reorganization rather than serving as simple proxies for tumor burden or patient age, thereby providing orthogonal information for survival modeling.

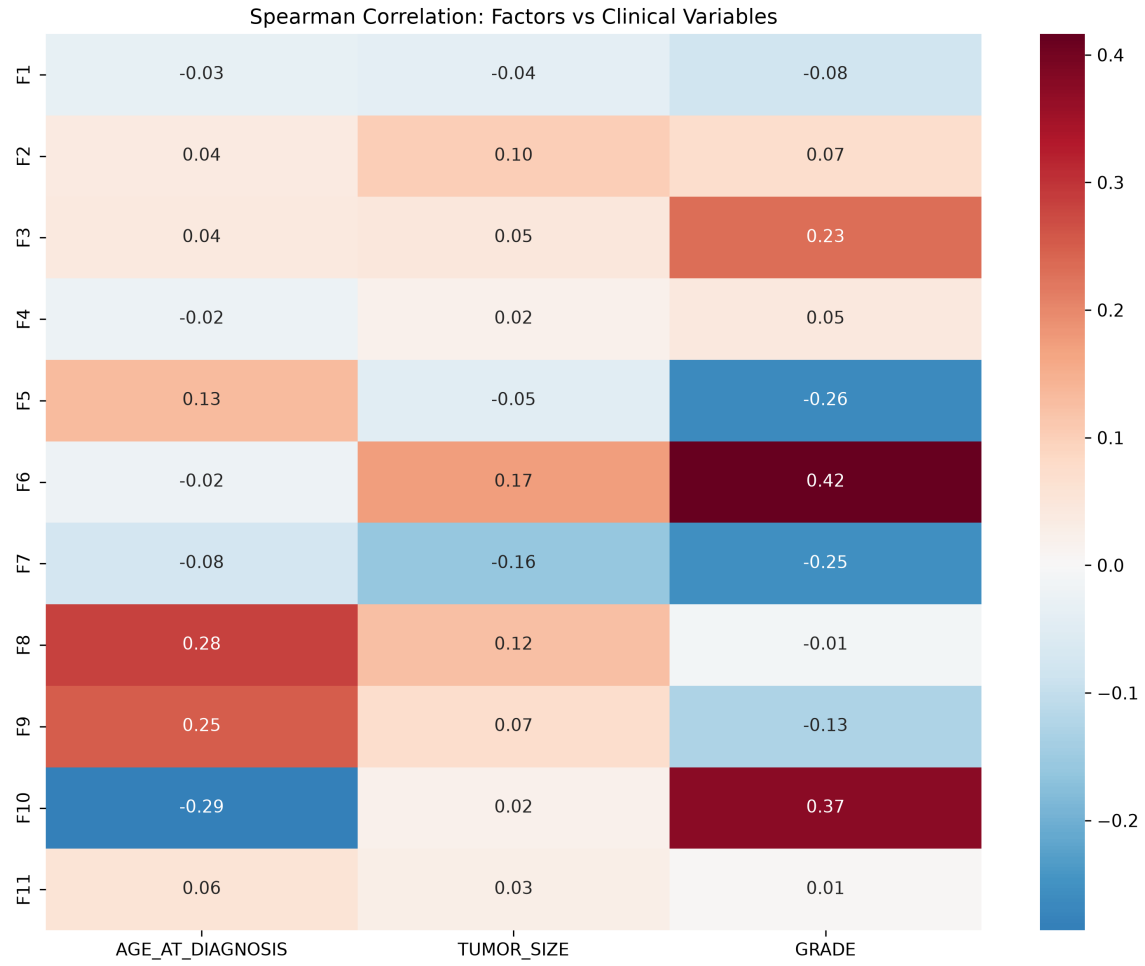


Figure 3: Heatmap displaying Spearman correlation coefficients between the 11 tensor-derived latent factors (F1–F11) and standard clinicopathological variables (`AGE_AT_DIAGNOSIS`, `TUMOR_SIZE`, and `GRADE`). The predominantly weak correlations indicate that these joint regulatory architectures capture intrinsic molecular reorganization rather than serving as simple proxies for patient age or tumor burden. This independence underscores the prognostic utility of the factors, which maintain their predictive significance for disease-specific mortality even when adjusted for these standard clinical risk factors.

3.3 Biological grounding of latent factors

To provide biological context for the latent factors, we performed a feature importance analysis on the gene-mode weights. The analysis revealed that the factors represent broad transcriptional and epigenetic

programs rather than being driven by isolated somatic mutations in known drivers. Across most factors, there was a lack of statistically significant enrichment for a panel of 173 known cancer driver genes (all driver overlap $p \geq 0.457$).

Specific factors were characterized by distinct biological themes. For instance, F6, which emerged as a high-risk state, was heavily weighted by proliferation and cell-cycle genes, including *TPX2* (rank 1), *CENPL*, and *MCM2*. Conversely, protective factors such as F3 and F5 were characterized by different gene sets; F3 was driven by genes including *AGAP2*, *ETNPPL*, and *AHNAK*, while F5 featured *PLEKHM2*, *TAF9B*, and *ACVRL1*. Furthermore, we observed that specific high-risk molecular states mapped coherently onto established molecular labels. For example, F10 scores were notably elevated in ER-/HER2- and HER2+ tumors compared to ER+/HER2- subtypes (), providing further biological grounding for its strong association with breast cancer lethality.

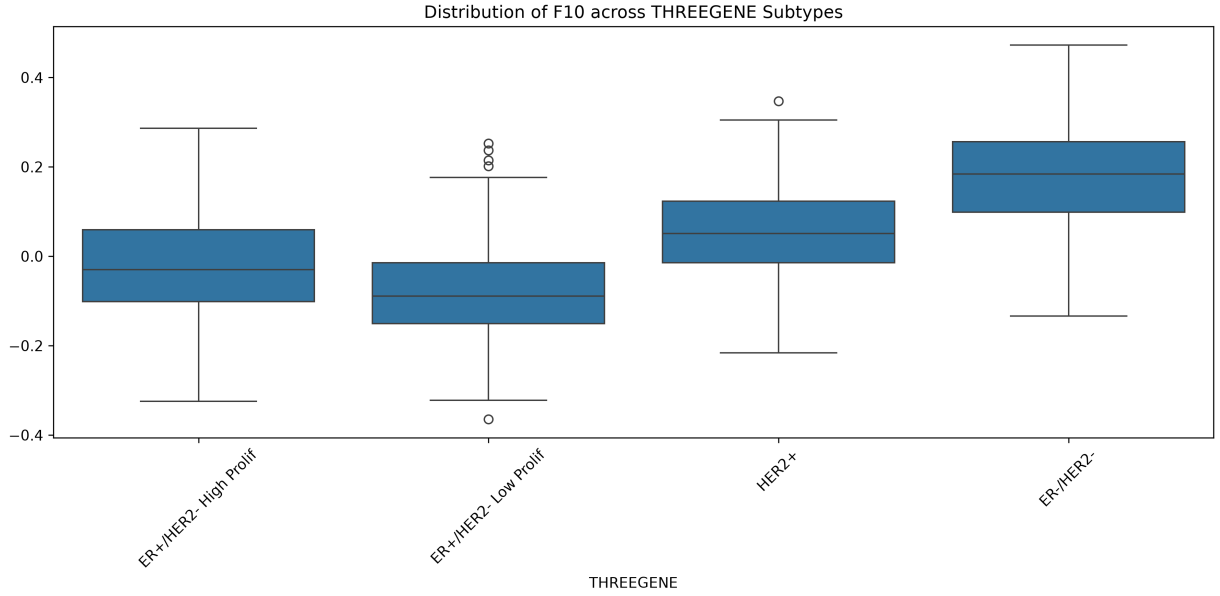


Figure 4: Boxplot illustrating the distribution of Factor 10 (F10) sample-mode scores across `THREEGENE` subtypes. F10, a latent factor derived from the multi-way tensor decomposition of three molecular layers, represents a high-risk joint regulatory architecture with a mean Hazard Ratio of 5.30 for disease-specific mortality across replication folds. The plot reveals that F10 scores are notably elevated in ER-/HER2- and HER2+ tumors compared to ER+/HER2- subtypes. This demonstrates how this specific high-risk molecular state maps onto established molecular labels, providing further biological context for its strong association with breast cancer lethality despite the overall independence of the tensor-derived factors from standard clinicopathological variables.

3.4 Predictive utility in competing risks and specificity to cancer lethality

The clinical impact of the joint regulatory states was evaluated using a cause-specific competing risks framework to explicitly isolate disease-specific mortality from other-cause deaths. In the Full Model—which adjusts for age at diagnosis, Integrative Clusters, histological grade, and tumor size—several factors emerged as powerful, independent drivers of disease-specific mortality. Notably, F6 represented a severe high-risk architectural state with a Hazard Ratio (HR) of 29.83 ($p = 0.0005$), and F8 demonstrated a significant deleterious effect with an HR of 2.57 ($p = 0.0014$). In contrast, F5 and F3 emerged as protective molecular states, with HRs of 0.06 ($p < 0.0001$) and 0.41 ($p = 0.0016$), respectively.

Crucially, these factors demonstrated high specificity to cancer-related outcomes. Stratification by F6 scores revealed a marked divergence in the 10-year cumulative incidence of disease-specific death (34.7% in the high-score group versus 20.8% in the low-score group). However, the 10-year cumulative incidence of

death from other causes was virtually identical between the two groups (13.4% versus 12.9%), confirming that F6 captures intrinsic tumor lethality rather than general host fitness ().

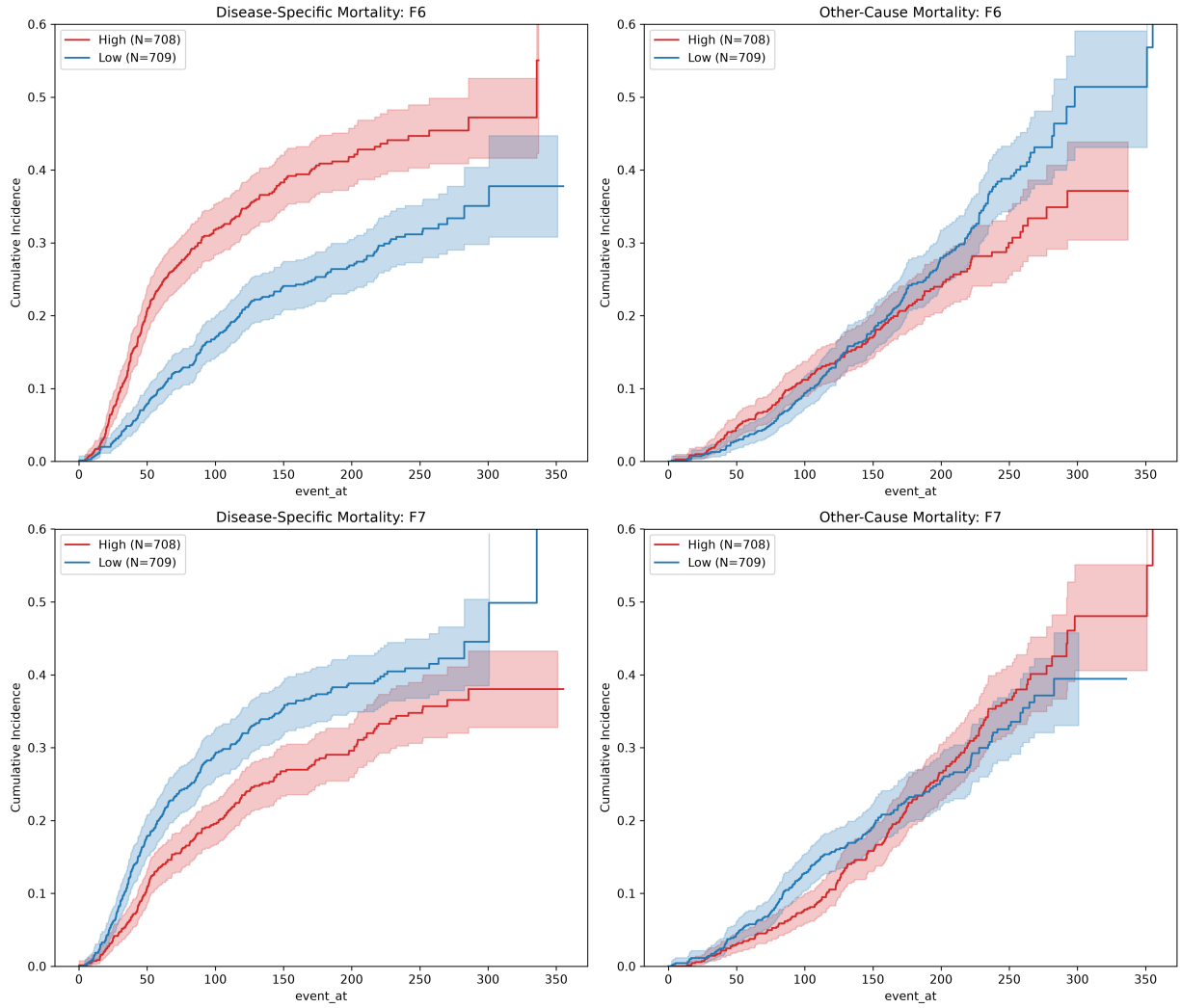


Figure 5: Cumulative incidence curves for disease-specific (left) and other-cause (right) mortality, stratified by high and low sample-mode scores for latent factors F6 (top) and F7 (bottom). The plots demonstrate the specificity of the tensor-derived joint regulatory architectures to cancer lethality. F6 emerges as a critical high-risk state, where high scores are associated with a markedly increased cumulative incidence of disease-specific death (34.7% versus 20.8% at 10 years). In stark contrast, F6 shows virtually identical other-cause mortality between the high (13.4%) and low (12.9%) groups. This differential impact, alongside the distinct stratification patterns observed for F7, confirms that these latent factors capture intrinsic molecular reorganisations specifically driving tumour lethality, rather than acting as proxies for general host fitness or aging.

The addition of the 11 latent factors to the Expanded Baseline Model substantially improved prognostic performance for disease-specific mortality. The Concordance Index (C-index) increased from 0.6834 to 0.7072, and the 10-year Brier score improved from 0.1861 to 0.1815, indicating enhanced predictive accuracy at the tails of the survival distribution while maintaining robust calibration (). In contrast, models predicting other-cause mortality showed only a marginal C-index change (0.7756 to 0.7788) and an increase in the Akaike Information Criterion (AIC) from 3790.3 to 3803.3 upon adding the latent factors, further underscoring their specificity to breast cancer lethality.

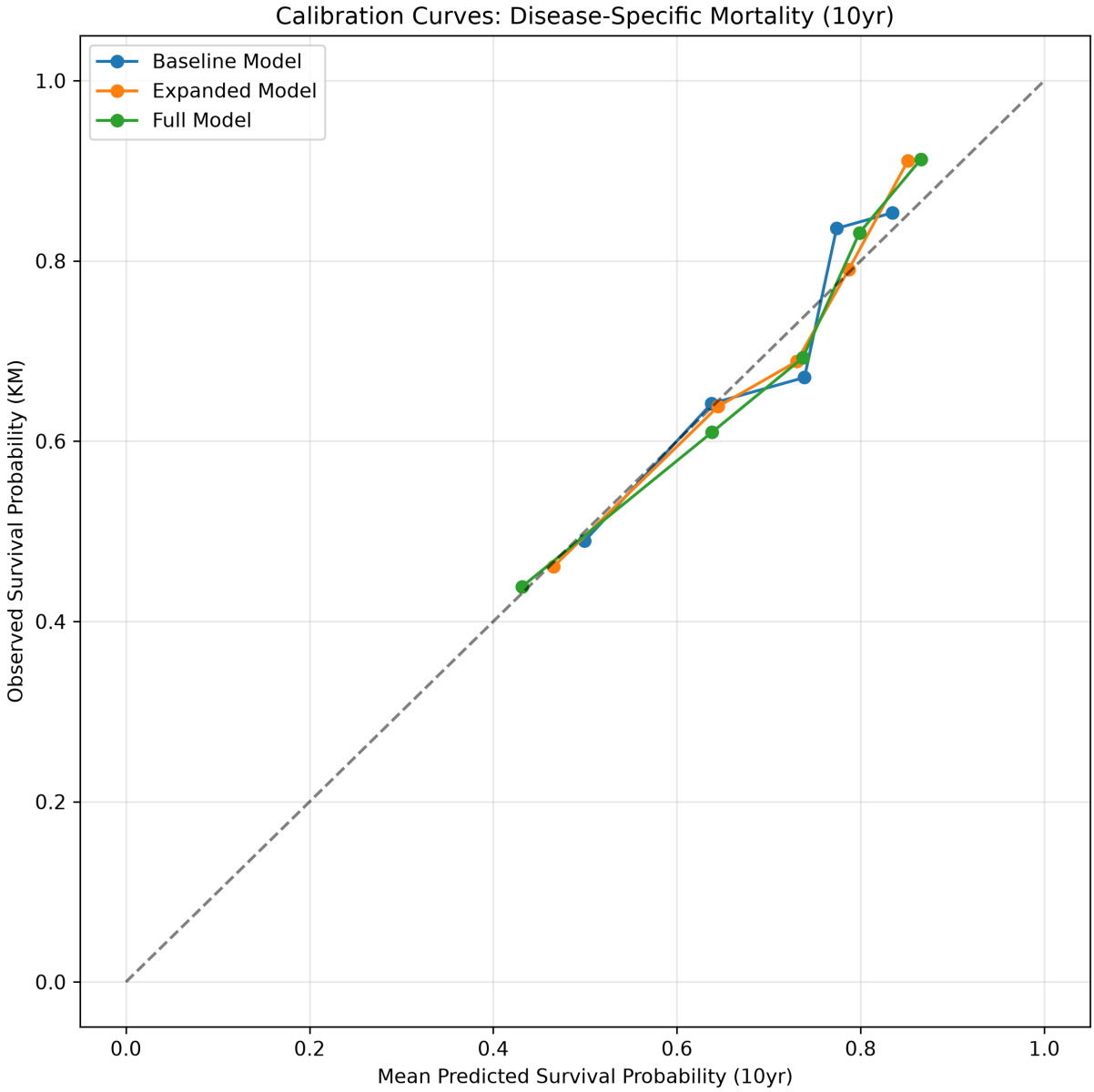


Figure 6: Calibration curves for 10-year disease-specific mortality comparing the Baseline, Expanded, and Full models. The Full Model integrates 11 tensor-derived latent factors with standard clinicopathological variables, including `AGE_AT_DIAGNOSIS`, `INTCLUST`, `GRADE`, and `TUMOR_SIZE`. The plot demonstrates that all models are well-calibrated, with mean predicted survival probabilities closely tracking the observed Kaplan-Meier (KM) estimates along the ideal dashed diagonal. Notably, the addition of the latent factors in the Full Model enhances overall predictive accuracy for breast cancer lethality, improving the Concordance Index from 0.6834 to 0.7072 and reducing the 10-year Brier score from 0.1861 to 0.1815, while maintaining robust calibration.

3.5 Accrual cohort replication

To validate the generalizability of the joint regulatory signals, we evaluated their predictive performance using a LOCO cross-validation strategy. The tensor decomposition revealed varying dependencies on specific omic layers across the factors, capturing both single-layer dominated signals and coupled multi-omic regulatory modes, such as joint dosage-driven or epigenetic-driven axes (, top-left).

When projected onto held-out cohorts, the latent factors maintained their prognostic relevance. The mean C-index for disease-specific mortality across the held-out cohorts was 0.6366 (SD = 0.018, range 0.6089 to 0.6602). Although slightly lower than the discovery C-index, the consistent performance above the 0.5 null across all five cohorts confirms the robustness of the joint regulatory signals (, top-right). Furthermore, while the direction and magnitude of the Hazard Ratios for individual factors exhibited some sensitivity to cohort composition, key factors remained strong predictors across folds; for instance, F10 demonstrated a mean HR of 5.30 across the replication folds (, bottom). This replication solidifies the utility of multi-way tensor decomposition in identifying reproducible molecular axes of breast cancer lethality.

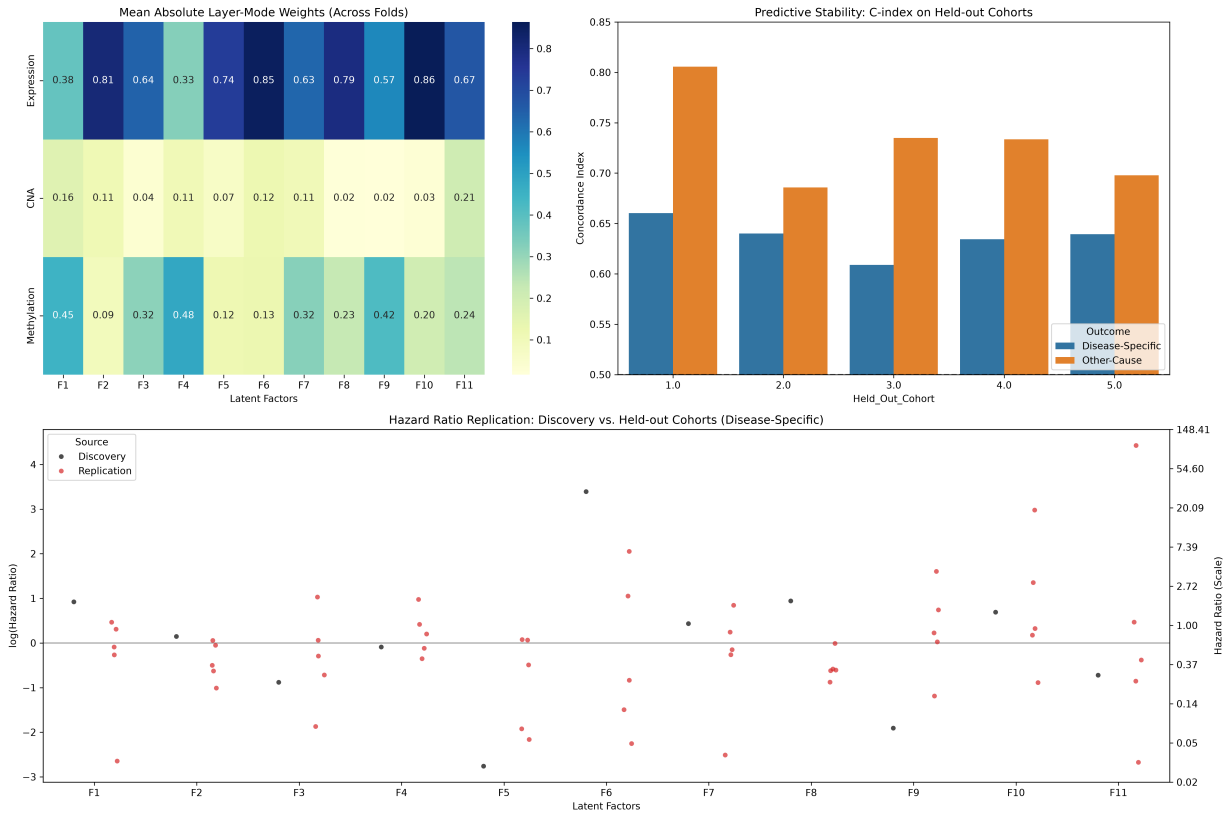


Figure 7: Multi-way tensor decomposition of joint regulatory architectures and their predictive stability across five accrual cohorts. The top-left panel displays the mean absolute layer-mode weights for the 11 latent factors, illustrating varying dependencies on gene expression, copy number alterations (CNA), and promoter methylation, which capture both single-layer dominated signals and coupled multi-omic regulatory modes. The top-right panel presents the Concordance Index (C-index) for disease-specific and other-cause mortality evaluated via a leave-one-cohort-out (LOCO) approach, confirming the generalisability of the joint regulatory signals with disease-specific predictive performance consistently remaining above the 0.5 null. The bottom panel compares the log hazard ratios of the latent factors between the discovery set and replication folds. This demonstrates that while these molecular axes serve as robust predictors of breast cancer lethality, the magnitude and direction of individual factor effects exhibit sensitivity to cohort composition.

4 Conclusions

In this study, we demonstrated that modeling the simultaneous interplay of genomic and epigenetic regulation through multi-way tensor decomposition provides a powerful and biologically grounded approach for predicting breast cancer lethality. By constructing a three-dimensional multi-omic tensor encompassing gene expression, copy number alterations, and promoter methylation across 1,417 primary tumors, we successfully extracted 11 structurally stable latent factors. Unlike traditional approaches that isolate molecular layers or rely on sequentially degraded models, these factors holistically encode joint cross-layer regulatory architectures, capturing both single-layer dominated signals and coupled multi-omic modes such as dosage-driven and epigenetic-driven axes.

A central achievement of this work is the explicit isolation of tumor-intrinsic lethality from general host fitness. By employing a cause-specific competing risks framework, we established that the prognostic utility of these latent molecular states is highly specific to disease-related mortality. The integration of these tensor-derived factors into survival models significantly improved the Concordance Index for disease-specific death from 0.6834 to 0.7072, while offering no predictive enhancement for other-cause mortality. Furthermore, these factors maintained their prognostic significance even after rigorous adjustment for standard clinicopathological variables, including age at diagnosis, tumor size, histological grade, and established molecular subtypes. This independence confirms that the identified regulatory architectures capture fundamental molecular reorganization rather than serving as redundant proxies for tumor burden or patient age.

Biological grounding of the latent factors revealed that breast cancer lethality is strongly associated with broad transcriptional and epigenetic programs rather than isolated somatic mutations in known cancer drivers. Specific configurations, such as the proliferation-enriched Factor 6, emerged as severe high-risk states that drastically increase the cumulative incidence of disease-specific death, whereas other configurations, such as Factors 3 and 5, conferred robust protective effects.

Finally, the structural stability and predictive generalizability of these joint regulatory signals were rigorously validated using a leave-one-cohort-out cross-validation strategy. The consistent performance of the gene-mode weights and prognostic models across five distinct accrual cohorts underscores the reproducibility of our findings. Ultimately, this multi-omic tensor framework overcomes the limitations of traditional single-layer analyses and confounded survival models, offering a robust methodology for stratifying breast cancer lethality that can be readily adapted to uncover complex, cross-layer drivers of progression in other malignancies.